Lecture 06: Hitchhiking and Selective Sweeps

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Overview

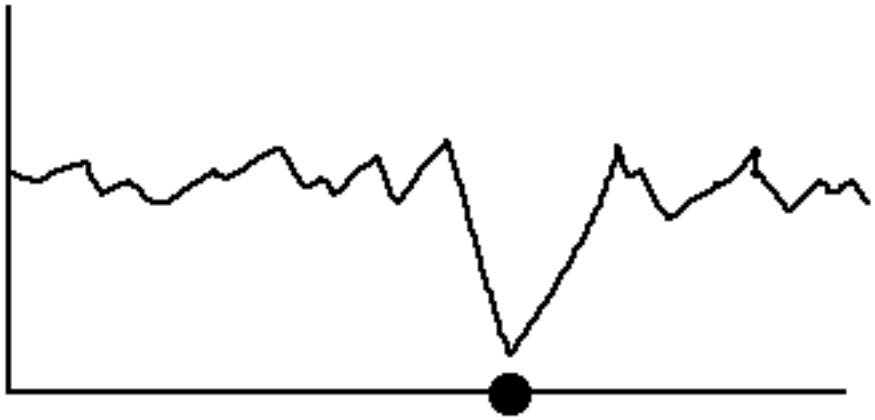
- Types of sweeps
- Impact on coalescent
- Hard vs. soft sweeps
- Population genetics of a sweep
- Standing sweeps and recurrent mutations
- Impact from recurrent sweeps
- Codon usage bias
- The Hill-Robertson effect
- All the details in WL Chapter 8

Hitchhiking

- When an allele is linked to a site under selection, its dynamics are considerably altered relative to drift
- A neutral mutation can hitchhike up to high frequencies when linked to a favorable mutation
- A sweep refers to the consequences of a recently-fixed site.
- A partial sweep occurs when a neutral allele has its frequency increased by a favorable allele increasing in the population
- A polygenic sweep is when an adaption occurs via small allele-frequency changes at a number of loci (little signal)

Sweeps and the coalescent

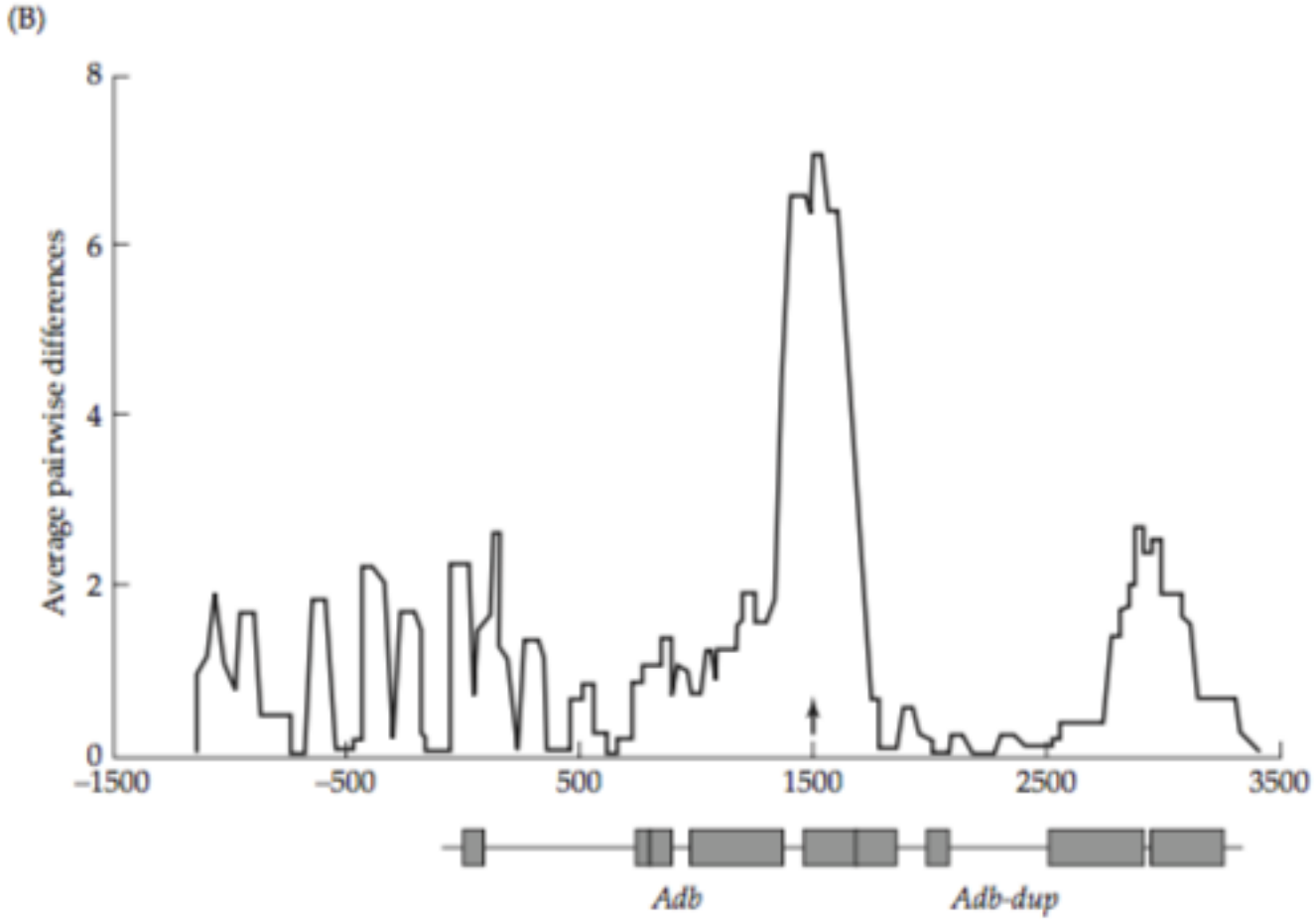
- A site linked to a recent sweep has a more recent TMRCA (Time to Most Recent Common Ancestor = coalescent time) relative to an unlinked neutral site
- Shorter TMRCA = less variation
  - Indeed, the term sweep refers to the "sweeping" away of linked neutral variation around a recently fixed site
- Under long-term balancing selection, longer time to MRCA relative to neutral sites = more variation



Site under
Directional
selection

Site under
Long-term
Balancing selection

Balancing selection = alleles favored when rare (overdominance, frequency-dependent selection)
Long term = greater than $4N_e$ generations

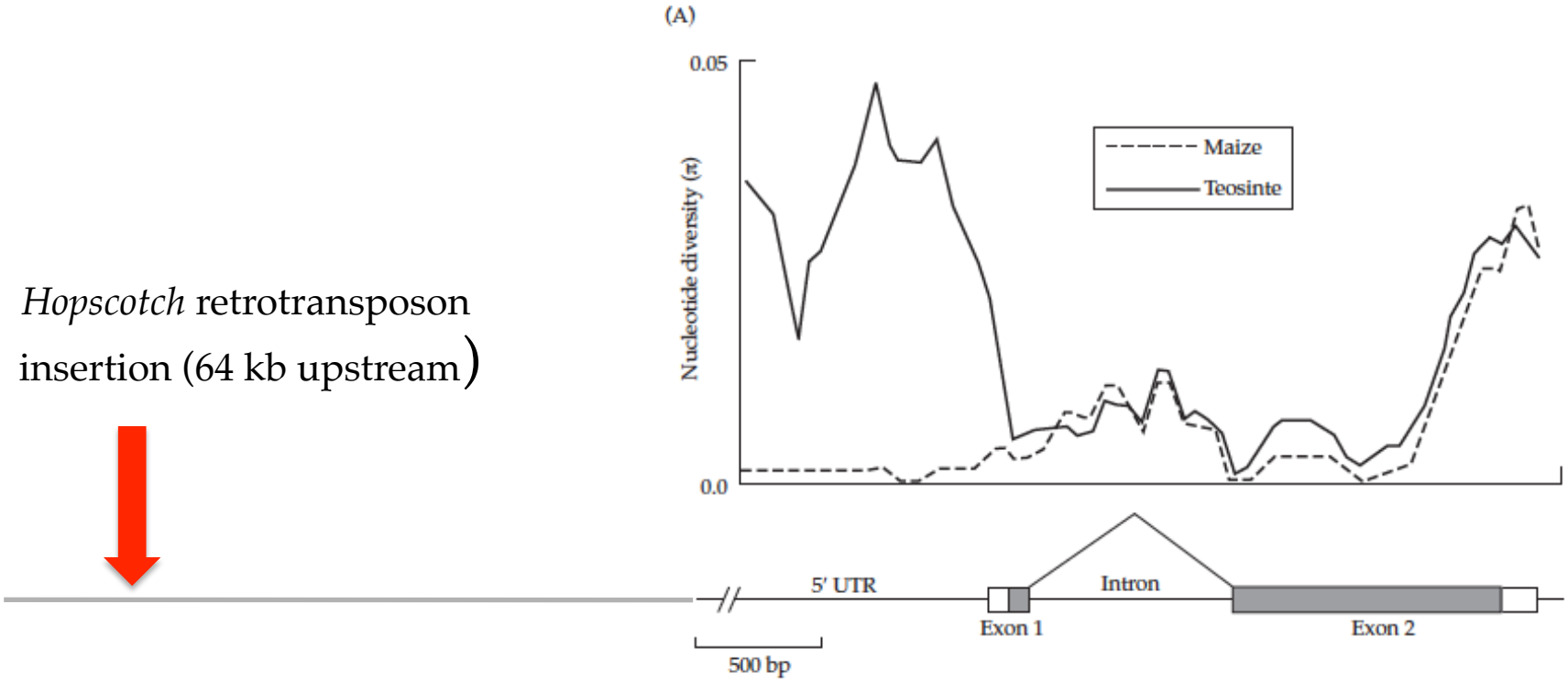


Example: Fast/slow allele of ADH in Drosophila shows signature of long-term balancing selection

tb1 is a key gene in this domestication

Domestication: Maize vs. teosinte

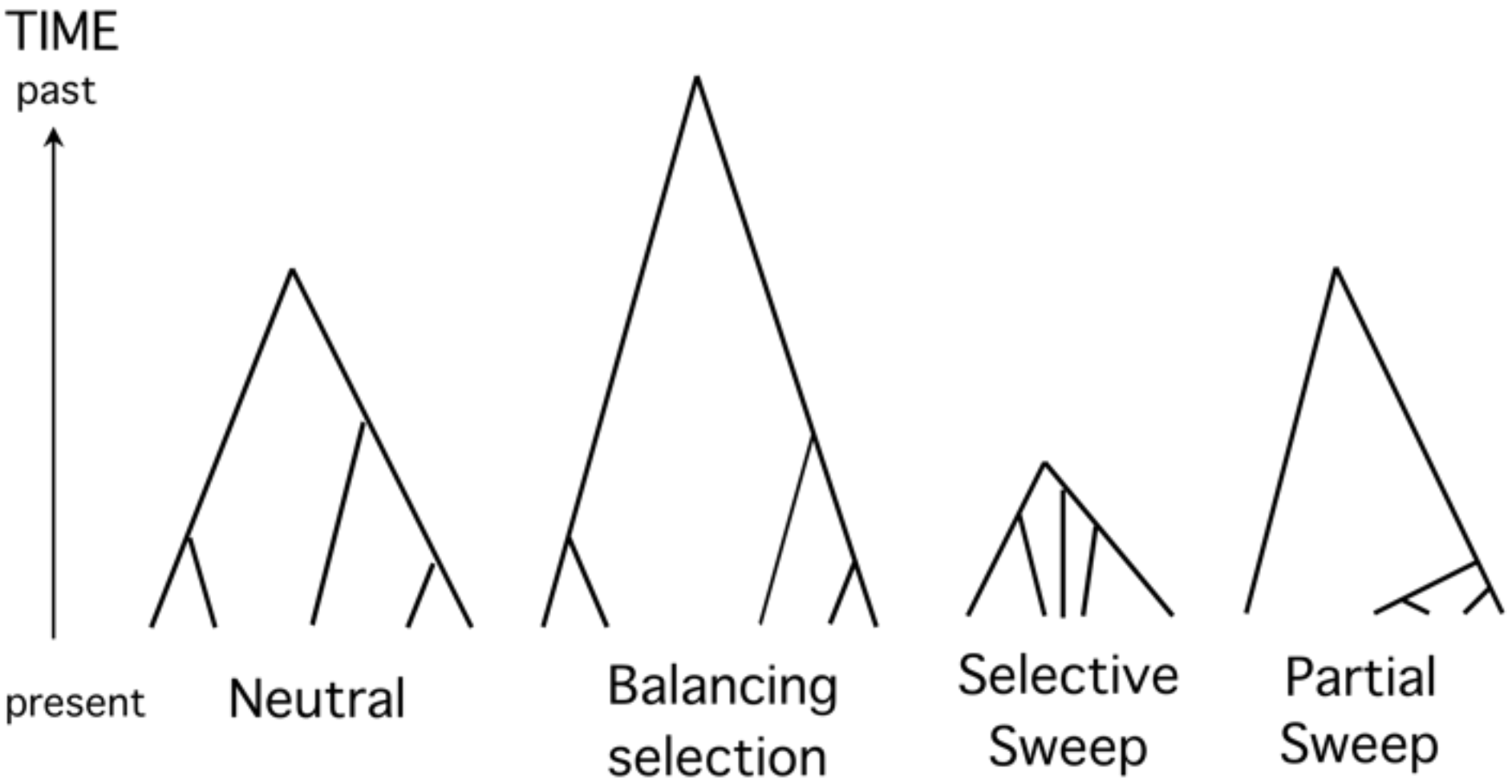
Hopscotch retrotransposon insertion (64 kb upstream)



Example: tb1 locus in maize has reduced variation in its 5′ region relative to its ancestor -- signal of a sweep, and likely a domestication gene

Selection changes the shape of the coalescent

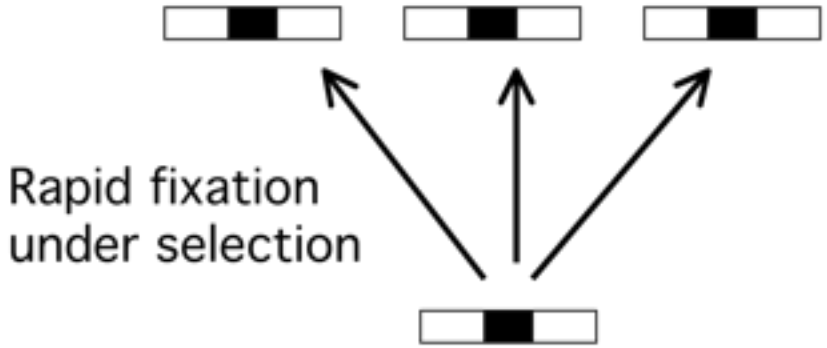
- A sweep not only changes the total size of a coalescent (by changing the TMRCA), it also changes its shape
- Under drift, nodal lengths i*ncrease* as one goes back in time ($t_2 > t_3$, etc)
- Under a sweep, nodes are *compressed* as we move back in time
  - Star genealogy -- all nodes essentially equal
- The structure under a partial sweep and long-term balancing selection also different from drift
- Changes in shape change the pattern of distribution of variation (more rare alleles, etc)


TIME
past
present
Neutral
Balancing
selection
Selective
Sweep
Partial
Sweep

Hard vs. soft sweeps

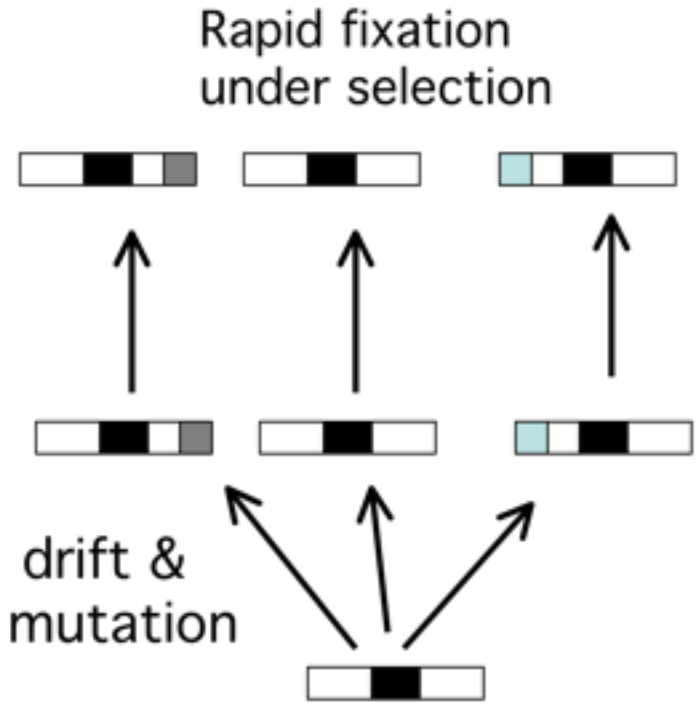
- A hard sweep is when a single new mutation arises and is immediately favored by selection --- drags along a single haplotype
- A soft-sweep is when either
  - A single mutation appears and then drifts around before it is favored (single-origin soft sweep)
  - Multiple mutations arise that (eventually) become advantageous (multiple-origins soft sweep)
  - Less signal with a hard sweep

A) Hard Sweep



Selection starts at the
appearance the new mutation

B) Soft Sweep



Initially, new mutation
neutral

Population genetics of sweeps

- Race between selection fixing a site and recombination removing initial associations
- Let f_s = fraction of initial association remaining after a sweep

Fraction f_s of initial associations remaining at fixation:

$$f_s \simeq \begin{cases} (p_0)^{-c/(2hs)} \simeq 1 - \dfrac{c}{2hs}\ln(p_0) & \text{for } p_0 \gg 1/(2N_e s) \\ (4N_e s)^{-c/(2hs)} \simeq 1 - \dfrac{c}{2hs}\ln(4N_e s) & \text{for } p_0 = 1/(2N) \end{cases}$$

Neutral allele frequency change

- Let neutral allele A be the allele linked to a favorable new mutation
- If q is the initial frequency of A
  - Total frequency change $\Delta q = (1-q)f_s$
  - Final allele frequency following sweep $q_h = q + \Delta q = f_s + q(1-f_s)$

Reduction in heterozygosity

- Rough rule (Kaplan & Hudson)
  - Sites within ~ 0.01s/c of a selected sites show significant reduction in H
  - Hence, if L is the length of reduction, then s ~ cL/0.02
  - Suppose a sweep covers 50kb (0.05MB) and c ~ 2 Cm/Mb, then s ~ 0.05*0.02/0.02 = 0.05
- More accurate value (additive favorable gene):
  - $H_h/H_0 \sim 1-(4NEs)^{-2c/s}$

For s = 0.01, $N_e = 10^6$, 1cM/Mb,

H_h/H_0

	1 kb	5kb	10 kb	25 kb	50 kb	100 kb
Dominant	0.01	0.05	0.10	0.23	0.41	0.65
Additive	0.02	0.10	0.19	0.41	0.65	0.88
Recessive	0.17	0.50	0.67	0.83	0.91	0.95

Exact expressions given in Chapter 8

Favorable recessive leave a very short signature of reduced heterozygosity

Recovery of Variation Following a Sweep

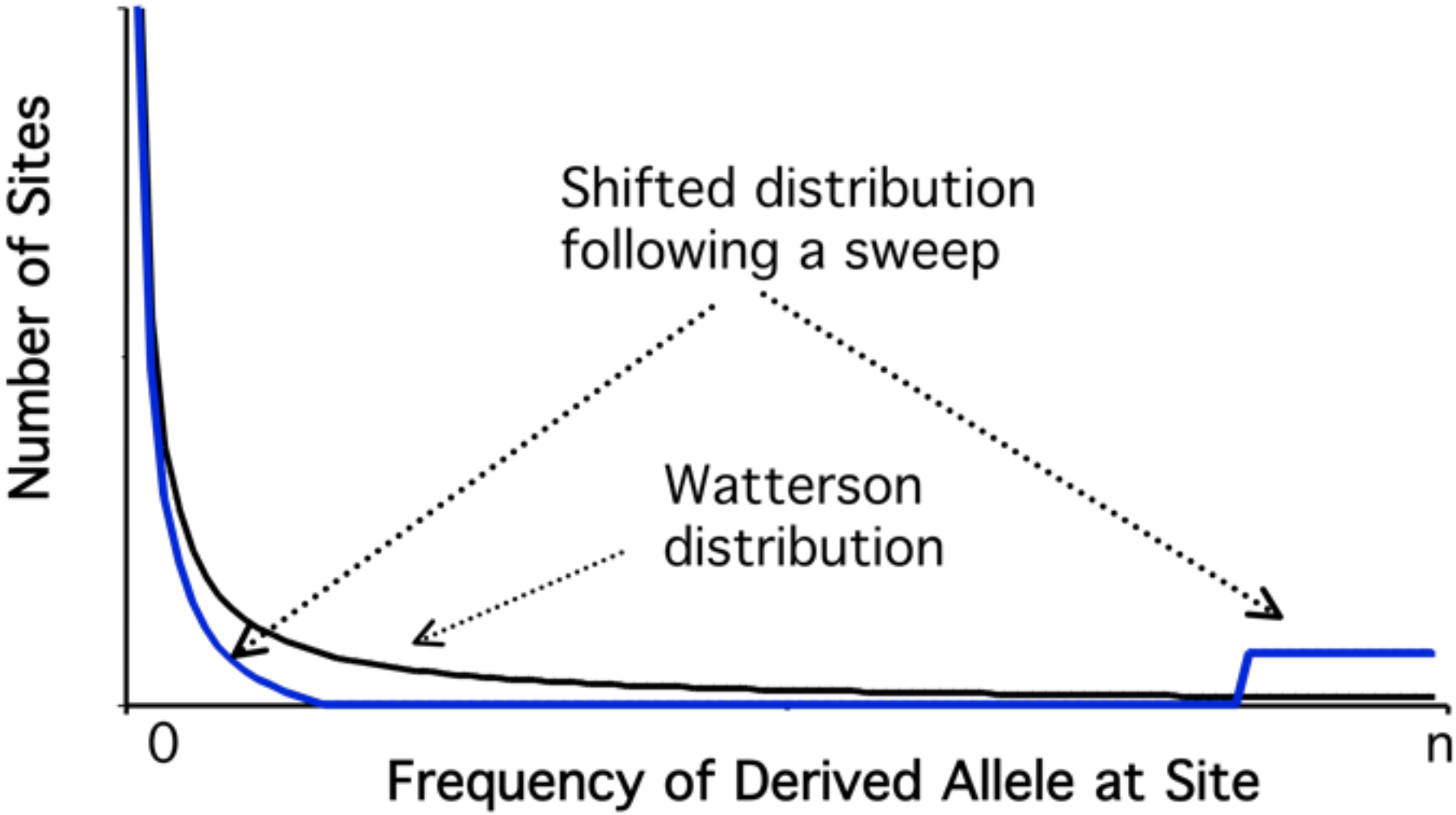
The signal left by even a strong sweep is a transient one, as new mutation will eventually restore heterozygosity at the neutral site back to its equilibrium value ($H_0 = 4N_e\mu$) before the sweep. Kim and Stephan (2000) find that the expected heterozygosity t generations after a sweep is approximately

$$E[\,H(t)\,] \simeq H_0 \left(1 - (4N_e s)^{-2c/s} \cdot e^{-t/(2N_e)}\right) \tag{7.11}$$

where $-H_0(4N_e s)^{-2c/s} = -H_0\, f_s$ is the reduction immediately following the sweep, which decays away by $1/(2N_e)$ each generation, as $(1 - 1/2N_e)^t \simeq \exp(-t/2N_e)$. The expected time to recover half the variation lost during the sweep (its half-life) is $\exp(-t_{0.5}/2N_e) = 0.5$ or $t_{0.5} = -2\ln(0.5)N_e \simeq 1.4N_e$. Note the important result that $E[\,H(t)\,]/H_0$ is *independent* of the actual mutation rate μ. The reason is that a low (or high) mutation rate means both a slow (or fast) accumulation of new mutations following the sweep, but a low (or high) target heterozygosity to reach.

The site-frequency spectrum

- Under the infinite-sites model, one can consider the number of sites with exactly k copies of the derived (mutated) allele
- Under mutation-drift equilibrium, given by the Watterson distribution, where n_i = number of sites with exactly i copies of the derived allele
  - $E(n_i) = \theta/i$
  - A sweep shifts this distribution

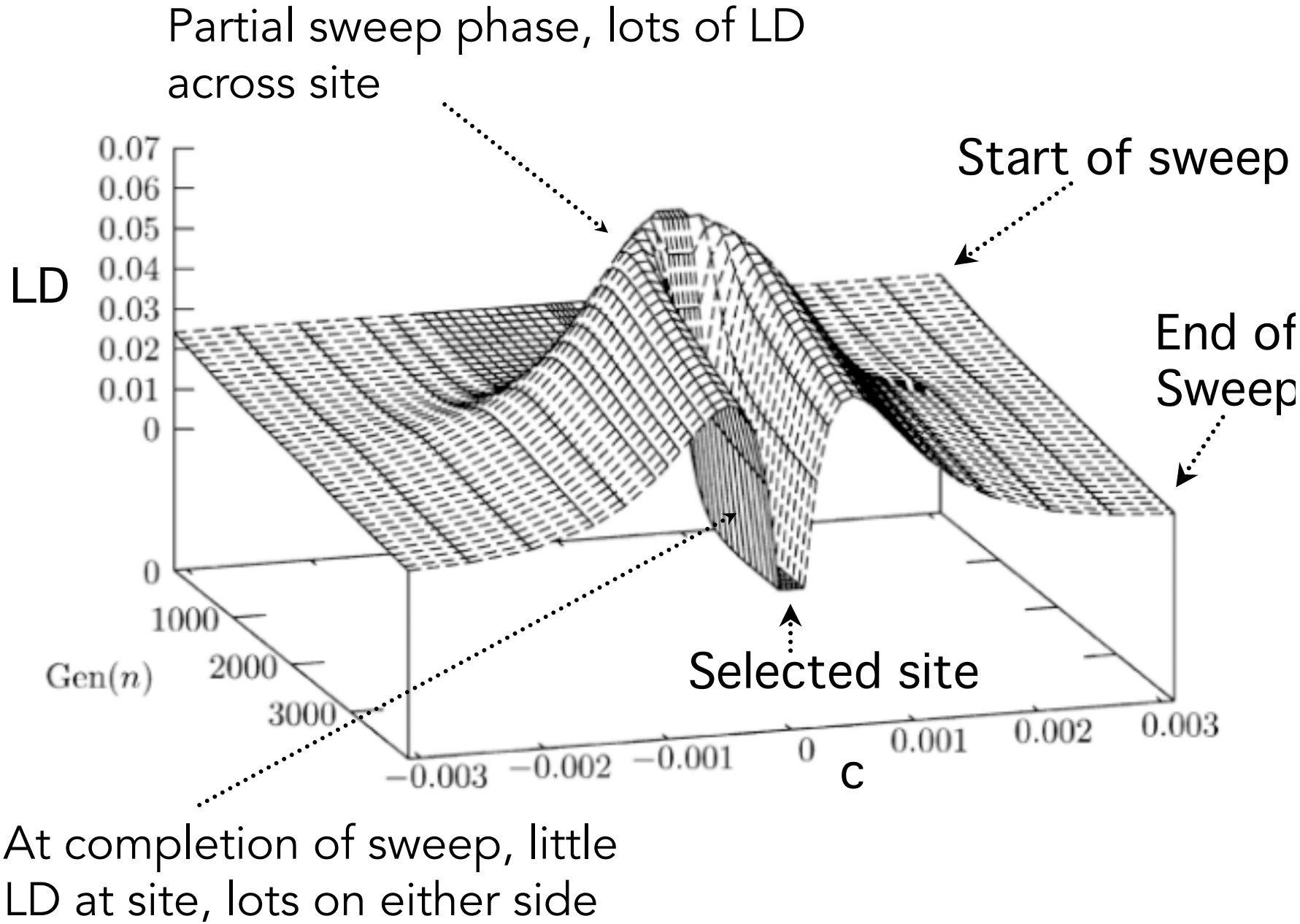


Generates (i) An excess of sites with high-frequency of derived alleles

(ii) An excess of sites with rare alleles

A sweep impacts LD around a site

- Initially, generates lots of LD across sites (partial sweep signature)
- However, as favorable allele fixed, little LD at selected site, but lots on either side of the site


Partial sweep phase, lots of LD across site
Start of sweep
End of Sweep
LD
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
1000
2000
3000
Gen(n)
Selected site
−0.003
−0.002
−0.001
0
0.001
0.002
0.003
c
At completion of sweep, little LD at site, lots on either side

Summary: Hard-sweep signal

A recent or ongoing sweep leaves several potentially diagnostic signals:

(1) *An excess of sites with rare alleles (in either the folded or unfold frequency spectrum)*

(2) *An excess of sites with high frequency* derived *alleles in the unfold frequency spectrum*

(3) *Depression of genetic variation, often* asymmetrically, *around the site of selection*

Signatures in the spatial pattern of LD differ during the sweep and after its completion:

When a favorable allele is at moderate frequencies (a partial sweep), we see

(4a) *An excess in LD throughout the region surrounding the sweep*

Following fixation of the favorable allele, the spatial pattern is rather different,

(4b) *An excess in LD on either side of the site, but a* depression *in LD around the site*

Finally,

(5) *Signatures of a sweep are very fleeting, remaining on the order of* $0.5N_e$ *generations for signature (1),* $0.4N_e$ *gens. for (2),* $1.4N_e$ *gens. for (3) and* $0.1N_e$ *gens. for (4b)*

Adaptation from standing variation

- Critical question: *How often does adaptation occur from pre-existing (standing) variation?*
- When the environment changes, can adaptation start right away or does it have to wait for new favorable mutation?
- Results from artificial selection experiments: lots of variation for just about any trait

Example 7.5. The threespine stickleback (*Gasterosteus aculeatus*) is a species (or species complex) of small fish widespread throughout the Northern Hemisphere in both freshwater and marine environments. The marine form is usually armored with a series of over 30 bony plates running the length of the body, while exclusively freshwater forms (which presumably arose from marine populations following the melting of the last glaciers) often lack some, or all, of these plates. Given the isolation of the freshwater lakes, it is clear that the reduced armor phenotype has independently evolved multiple times. Colosimo et al. (2005) showed that this parallel evolution occurred by repeated fixation of alleles at the *Eda* gene involved in the ectodysplasin signaling pathway. Surveying populations from Europe, North America, and Japan, they found that nuclear genes showed a clear Atlantic/Pacific diversion. Conversely, at the *Eda* gene, low armored populations shared a more recent history than full-armored populations, independent of their geographic origins, presumably reflecting more recent ancestry at the site due to the sharing a common allele. In marine populations, low-armored alleles at *Eda* are present a low (less than five percent) frequency. Presumably, these existing alleles were repeatedly selected following the colonization of freshwater lakes from marine founder populations.

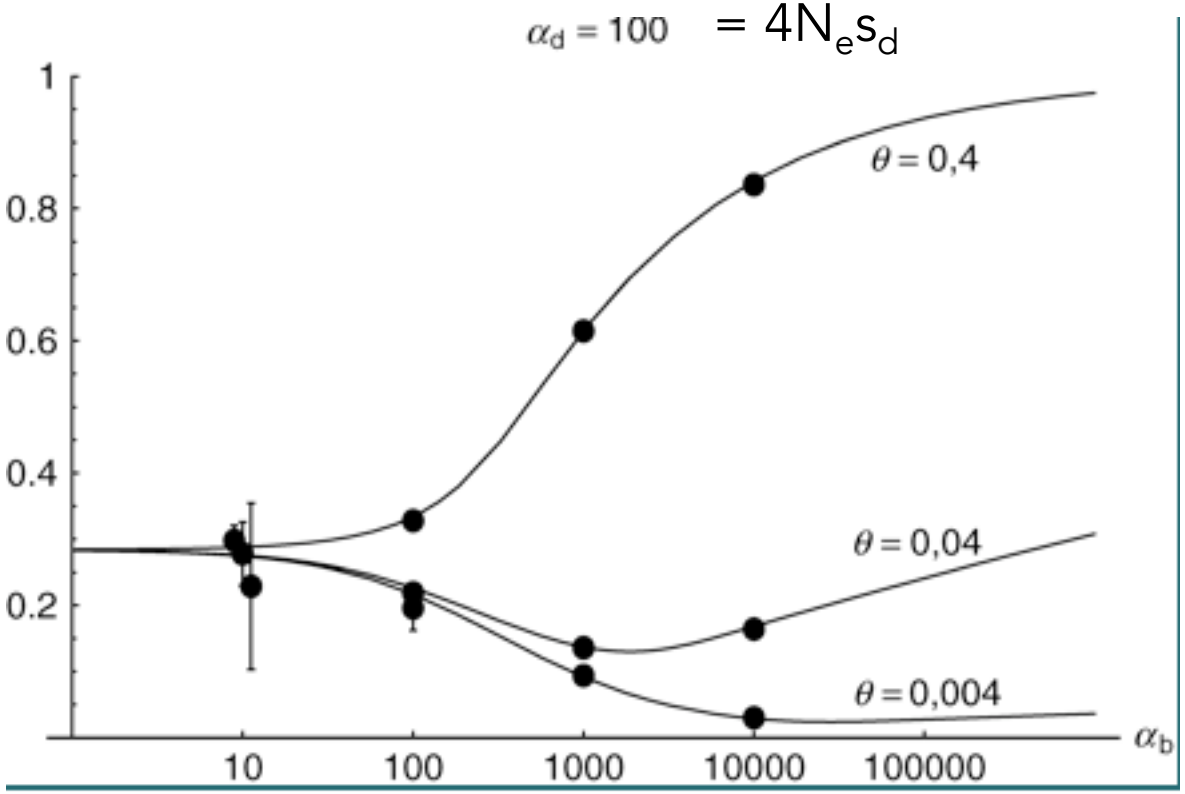
How likely is a sweep using standing variation?

- Hermisson & Pennings (2005) assumed an allele has fitness 1: 1-2$h_d s_d$: 1-2s_d in the old environment and 1: 1+2hs: 1+2s in the new.
- Probability that an existing allele is fixed becomes

$$\Pr_{sv} \approx 1 - \exp\left[-\theta_b \ln(1+R)\right], \quad \text{where} \quad R = \frac{2h\alpha_b}{2h_d\alpha_d + 1} \tag{7.22b}$$

with $\alpha_b = 4N_e s$ and $\alpha_d = 4N_e s_d$ are the scaled strenghts of selection in the new and old environments, respectively, and $\theta_b = 4N_e \mu_b$ the scaled benefical mutation rate.

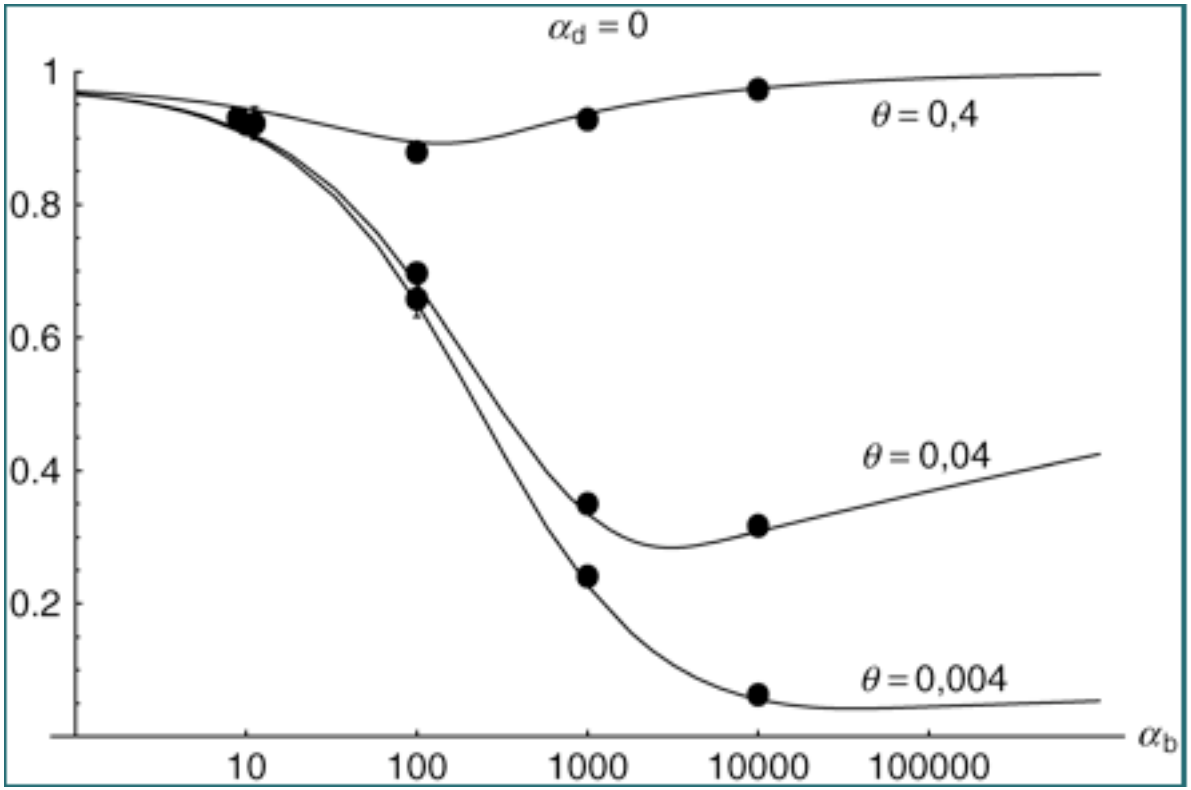
Prob(sweep from standing variation)



$\alpha_b = 4N_e s$

If allele is unfavorable in old environment, both a strong favorable effect in new ($\alpha_b >> 1$) and a high mutation rate are required

Prob(sweep from standing variation)



$$\alpha_b = 4N_e s$$

If allele is neutral in old environment, a sweep from Standing variation likely is mutation rate is modest

Example 7.6. Suppose $N_e = 10^6$ and the per-site mutation rate throughout the genome is $\theta = 0.01$. For a beneficial mutation that can only occur by a change to a specific nucleotide at a specific site, 1/3 of mutations at that site are beneficial, giving $\theta_b = 0.0033$. For an additive allele ($h = 1/2$) with $s = 10^{-4}$, we have $\alpha_b = 4 \cdot 10^6 \cdot 10^{-4} = 400$. If this mutation was neutral before being favored, $\alpha_d = 0$, $R = 2h\alpha_b = 400$ and Equation 7.22b gives

$$\Pr_{sv} \approx 1 - \exp\left[-\theta_b \ln(1+R)\right] = 1 - \exp\left[-0.0033 \ln(1+400)\right] = 0.013$$

Hence, there is only a once percent chance that a sweep occurs at this locus in the absence of new mutation. Now suppose that we examine this population at $T = 0.5$ (N_e generations). The probability that at least one such mutation destined to become fixed arises by this time is

$$\Pr_{new}(T) = 1 - \exp\left(-Th\alpha_b\theta_b\right) = 1 - \exp\left[-0.5 \cdot (1/2) \cdot 400 \cdot 0.0033\right] = 0.281$$

Provided we see a sweep at this locus by N_e generations, the probability it was due to an existing allele present at the time the environment shifted is

$$\pi_{ex} = \frac{\Pr_{sv}}{\Pr_{sv} + (1 - \Pr_{sv})\Pr_{new}(T)} = \frac{0.013}{0.013 + (1 - 0.013)0.281} = 0.05$$

giving only a five percent chance that the fixed favorable allele was present in the population at the start of selection.

Recurrent mutation

- Another possibility is that the new favorable mutation can arise several times during selection.
- If so, this means that selection fixes a set consisting of multiple haplotypes, leaving a very weak signal
- This is called a multiple-origins soft sweep
- How likely? Key is θ_b = scaled beneficial mutation rate
  - $\theta_b < 0.01$ very rare,
  - $0.01 < \theta_b < 1$ intermediate,
  - $\theta_b > 1$ almost certain

Example 7.9. Karasov et al. (2010) examined *Drosophila melanogaster* mutations at the *Ace* gene, which codes for the neural signaling enzyme Acetylcholinesterae, a target for many commonly used insecticides. Single nucleotide changes at four highly conserved sites confer partial insecticide resistance, with combinations of these conferring significantly greater resistance. Single, double, and triple mutations are all found in natural populations. While one model is that these variants existed at the start of major insecticide use (the 1950's), the authors found that mutations in North American and Australia appeared to have arise *de novo* following the *melanogaster* migration out of Africa. Given that only 1000 to 1500 generations have elapsed since the widespread use of insecticides that target the *Ace* product, estimates of $\theta \sim 0.01$ based on nucleotide diversity (and hence a θ_b of 1/3 this value at each of the four sites) are not consistent with the independent origins of single, much less multiple, mutations in this gene over this short time scale. However, if the actual effective population size was 10^8 instead of the standard assumed value of 10^6 during the past 50 years, then $\theta_b \sim 1$, and such multiple independent origins are highly likely. The effective population size that matters for these mutations is that during their origin and spread, not that set by any history predating their appearance.

Signatures from a soft-sweep

- Very little reduction in H is quite possible under a soft sweep, so H is not a good signal
- Good LD signal: under a soft-sweep, LD extends through the site of selection

Polygenic sweeps

- What if most adaptation occurs via genes of small effect? For a gene with s = 0.001, a sweep influences roughly 2000 bases for c = 1 cM / Mb
  - This is the best case (hard sweep)
- More generally, sweep could occur from relatively small allele frequency changes over a large number of loci -- no classic signature
- However, might find correlation in allele frequencies when different populations sampled over similar environments
  - Coop found this for some human genes

Genome-wide impact of recurrent selection

- Recurrent sweeps (RS) vs. background selection (BGS)
- Charlesworth: Background selection is the removal of deleterious mutations, which will reduce N_e at linked sites.
- Very hard to distinguish BGS from RS
- Common feature is that polymorphism is reduced in regions of low recombination
  - Both RS and BGS can explain this

Recurrent sweeps

For a population of constant size undergoing periodic sweeps, Wiehe and Stephan (1993) found that the equilibrium level of heterozygosity, measured by nucleotide diversity π, at linked neutral sites is approximately

$$\frac{\pi}{\pi_0} \simeq \frac{\rho}{\rho + \lambda\gamma\, k} \tag{7.29a}$$

where $\pi_0 = 4N_e\mu$ is the average heterozygosity at a single site for an equilibrium neutral population under no sweeps, ρ is the per-nucleotide recombination rate over the region of interest, $\gamma = 2N_e s$ the scaled strength of selection, λ the per-nucleotide adaptive substitution rate, and the constant $k \simeq 0.075$. Equation 7.29a assumes all new adaptive mutations have the same selective advantage. For modest values of ρ (relative to $\lambda\gamma k$), Equation 7.29a is approximately

$$\frac{\pi}{\pi_0} \simeq 1 - \frac{\lambda\gamma k}{\rho} \tag{7.29b}$$

A few large, or many small, sweeps

- The reduction in diversity is a function of $\lambda\gamma$, the product of the rate and strength of a sweep. The same reduction could be caused by a few large sweeps or many small sweeps.
- Chapters 8 and 10 discusses some methods to try to estimate these separately
  - One approach to estimate γ is the regression of nucleotide diversity (heterozygosity) on amino acid divergence (details on pp 247-248). Slope estimates γ.

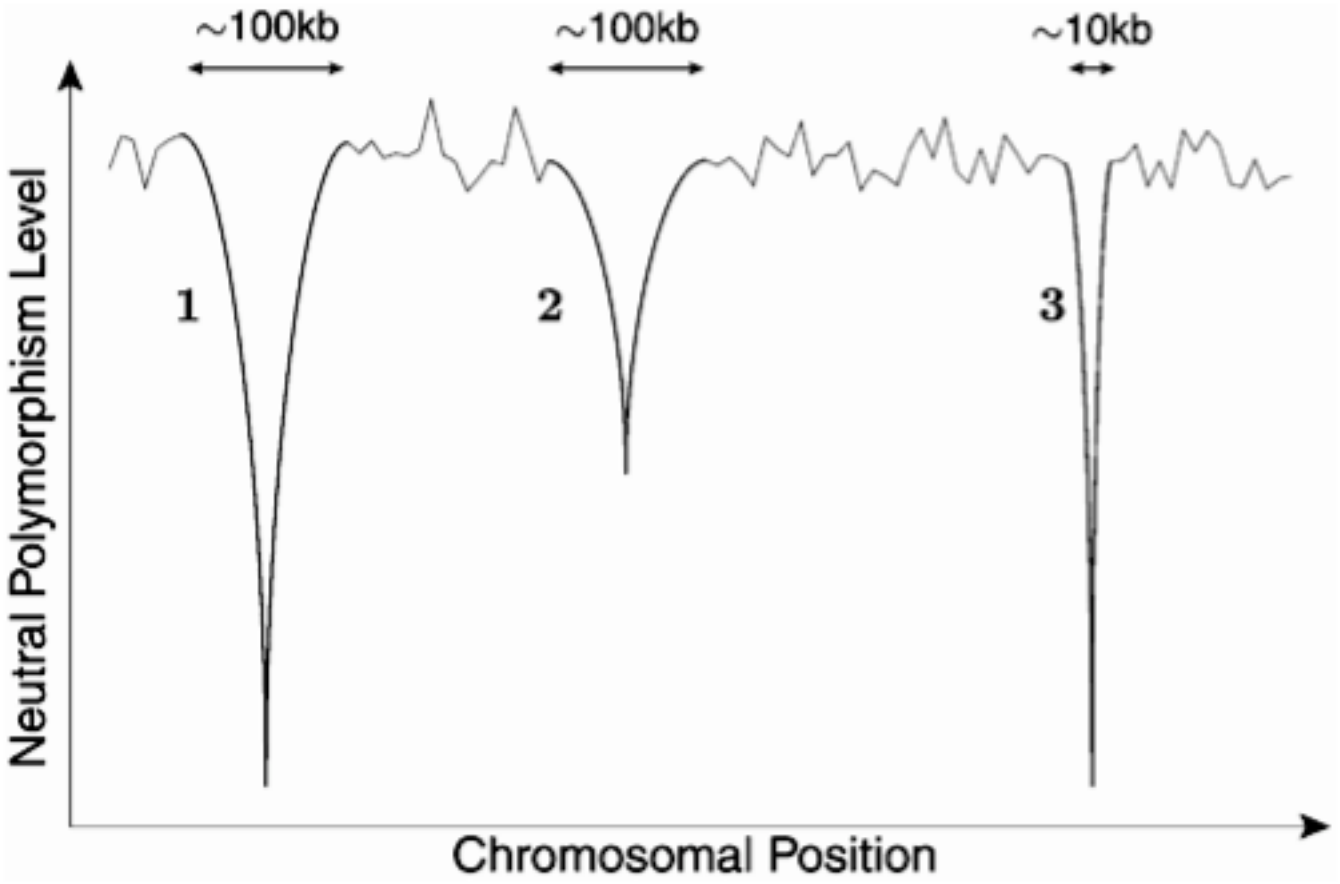
Table 7.3. Estimates of the rates of adaptive evolution at the molecular level for several *Drosophila* species and for the aspen tree (*Populus tremula*). The species listed provided the polymorphism data, while an outgroup was used for some estimates of λ (Equation 9.11a). Methods for estimating individuals components of the product $\lambda\gamma$ (the scaled strength of selection $\gamma = 2N_e s$, the rate of adaptive substitutions per base pair per generation λ, and the average strength of selection of a beneficial mutation s) are more fully developed in Chapter 9.

Organism	$\lambda\gamma$	γ	s	λ	Reference
D. melanogaster	3.9×10^{-7}	34,400	2.0×10^{-3}	6.0×10^{-11}	Li and Stephan 2006
D. melanogaster	5.1×10^{-8}	74	2.3×10^{-5}	7.0×10^{-10}	Bachtrog 2008
D. melanogaster	2.6×10^{-8}	35	1.2×10^{-5}	7.5×10^{-10}	Andolfatto 2007
D. melanogaster	4.0×10^{-7}	10,000	2.0×10^{-3}	4.2×10^{-11}	Jensen et al. 2008
D. simulans	1.1×10^{-7}	30,000	1.0×10^{-2}	3.6×10^{-12}	Macpherson et al. 2007
D. miranda	1.2×10^{-6}	3,100	2.7×10^{-3}	4.0×10^{-10}	Bachtrog 2008
D. melanogaster				1.8×10^{-11}	Smith & Eyre-Walker 2002
D. melanogaster				3.6×10^{-11}	Andolfatto 2005
D. melanogaster	1.3×10^{-8}				Wiehe & Stephan 1993
P. tremula	1.5×10^{-7}				Ingvarsson 2010
Humans				2.3×10^{-12}	Example 9.12

Example 7.11. As summarized in Table 7.3, for a set of X-linked genes in *D. melanogaster*, Andolfatto (2007) and Jensen et al. (2008) obtained estimates for λ of 7.5×10^{-10} and 4.2×10^{-11} (respectively). Consider a region of length 100 kb. Under Andolfatto's estimate, the per generation rate of adaptive substitutions over a region of this size is $10^5 \cdot 7.5 \times 10^{-10} = 7.5 \times 10^{-5}$ or one sweep roughly every 13,300 generations. Under Jensen's estimate, a sweep influencing this region occurs roughly every 238,000 generations.

Sweeps:
Many weak vs. few strong

- Another approach is the spatial pattern of variation, which should be different under a few strong vs. many weak
- Chapter 8 details approaches using this idea


~100kb
~100kb
~10kb
Neutral Polymorphism Level
1
2
3
Chromosomal Position

Sweeps, background selection & substitution rates

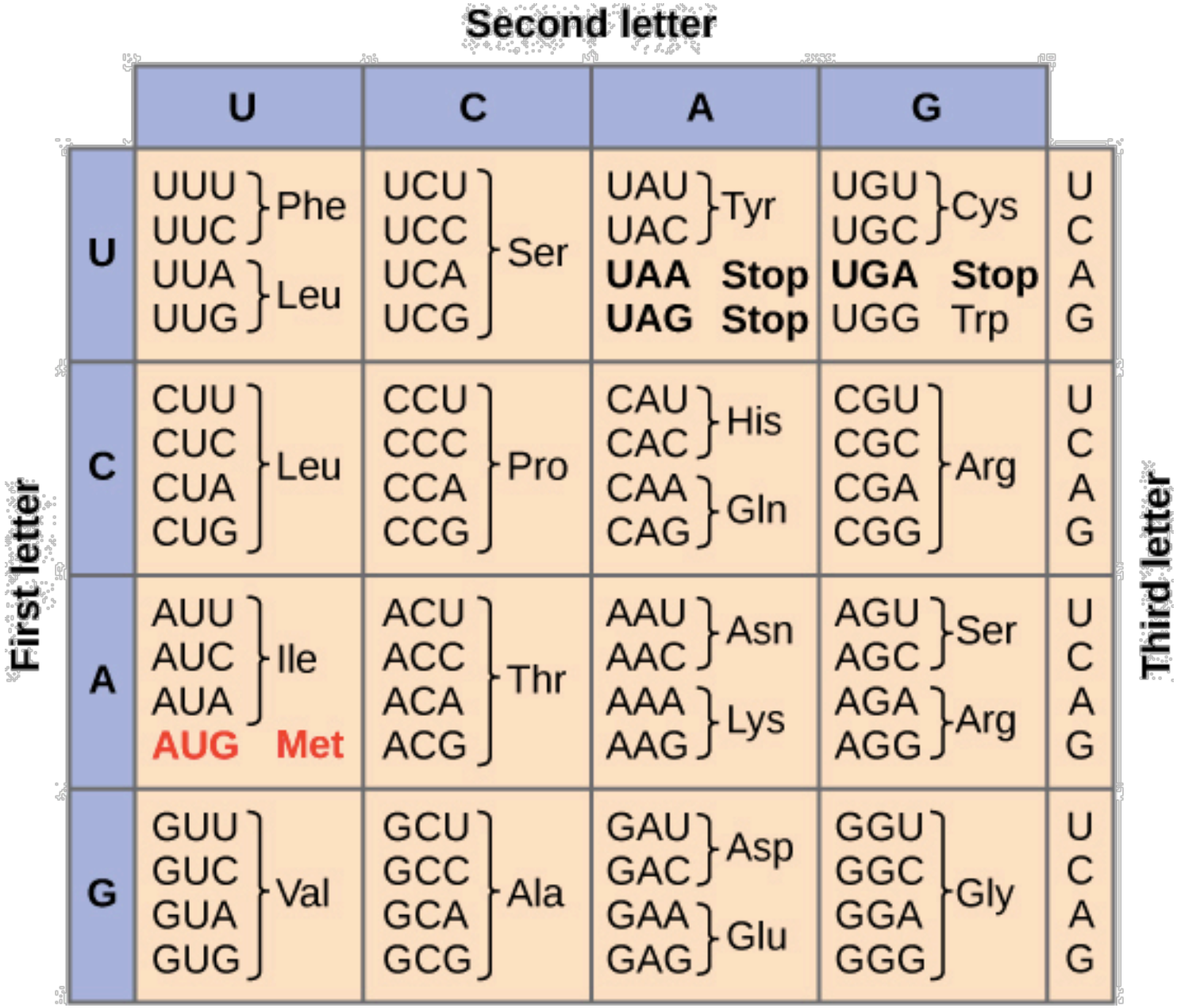
- A decrease in the effective population size should lead to
  - A decrease in the amount of polymorphism
  - An increase in the substitution rate (as more mutations become effectively neutral)
  - Further, since both RS and BGS should have a bigger impact in regions with lower recombination, what patterns are seen in the genome?

Substitution rates and recombination fraction

- *Drosophila*:
  - *D. melanogaster* vs. *D. yakuba* show an increase in the synonymous substitution rate on chromosomes with no recombination
  - Comparison of "dot" chromsome (no recombination) with autosomes: higher replacement substitution rates
- Human/chimp
  - No obvious effect of recombination on substitution rates

Correlation of divergence rate and polymorphism

- Sites with higher divergence rates for replacement sites may experience more sweeps, and hence have lower N_e and polymorphism
- Such a negative correlation between replacement divergence rate and level of synonymous site polymorphism seen in several species of Drosophila, European aspen, and humans
- Under BGS, sites with low divergence should also have lower levels of polymorphism (stronger constraints = lower divergence = more BGS)
  - Such a pattern seen in humans
  - However, could also arise from a simple reduction in mutation rate


Second letter
First letter
Third letter
U
C
A
G
UUU UUC Phe
UUA UUG Leu
UCU UCC UCA UCG Ser
UAU UAC Tyr
UAA Stop
UAG Stop
UGU UGC Cys
UGA Stop
UGG Trp
CUU CUC CUA CUG Leu
CCU CCC CCA CCG Pro
CAU CAC His
CAA CAG Gln
CGU CGC CGA CGG Arg
AUU AUC AUA Ile
AUG Met
ACU ACC ACA ACG Thr
AAU AAC Asn
AAA AAG Lys
AGU AGC Ser
AGA AGG Arg
GUU GUC GUA GUG Val
GCU GCC GCA GCG Ala
GAU GAC Asp
GAA GAG Glu
GGU GGC GGA GGG Gly

Codon usage bias

- Nonrandom use of synonymous codons is common in many species, and thought to result from selection more efficient/rapid translation by picking those codons corresponding to the most common tRNAs for that protein
  - Optimal (preferred or P) codons: should be (weakly) selected for
  - U = unpreferred codons should be weakly selected against
- Since s is expected to be small, modest changes in N_e may have a big signal, transforming a site under selection to a site that is effectively neutral
- By using an outgroup, Akaski compaired P -> U (mutations from preferred to unpreferred codons) with U -> P mutations in Drosophila
  - Excess of U -> P replacements in D. pseudoobscura
  - Excess of P -> U replacements in D. melanogaster (much smaller N_e)
  - P -> U mutations segregating at lower frequencies

Example 7.13. A related study was by Maside et al. (2004), who examined codon usage in *D. americana*, a member of the *virilis* species group. Using *virilis* as an outgroup, they observed 84 synonymous substitutions (fixed differences or divergence) between the two species and 144 segregating synonymous sites within *americana*. Classifying these as either a $P \to U$ or $U \to P$ showed the following pattern:

	Substitutions	Polymorphic (*americana*)	Polymorphism / Divergence
$P \to U$	52	124	2.38
$U \to P$	32	20	0.62

Fisher's exact tests gives $p = 6.4 \times 10^{-5}$, showing a highly significant deviation, with an almost four-fold higher polymorphism to divergence ratio for the putative deleterious mutations $P \to U$. Further, if this class is indeed deleterious, we would expect these mutations to be at lower frequencies in the sample than $U \to P$ mutations, and such a significant difference was observed. This difference in the site-frequency spectrum was first noticed by Akashi (1999) for *D. simulans*, which was shifted towards lower frequencies for unpreferred mutations and towards higher frequencies for preferred mutations.

Strength of selection on P codons

Given the above evidence for selection against unpreferred codons, how strong is selection? Using the Poisson random field (PRF) method for analysis of the pattern of fixed differences and polymorphic site (examined in detail in Chapter 9), estimates of $N_e|s| \sim 1$ were obtained for *simulans* and *pseudoobscura* (Akashi 1995, Akaski and Schaeffer 1997). An alternative approach to estimate $N_e|s|$ follows from Equation 6.35, which gives Li's (1987) expression for the expected frequency $\tilde{p}$ of a preferred allele at the mutation-selection-drift equilibrium. In the notation of this chapter, this becomes

$$\tilde{p} \simeq \frac{\exp(2\gamma)}{\exp(2\gamma) + \zeta} \tag{7.34}$$

where $\gamma = 2N_e s$ is the scaled strength of selection for preferred codons, and $\zeta = \mu_{P \to U}/\mu_{U \to P}$ measures any mutation bias (also see Bulmer 1991; McVean and Charlesworth 1999, 2000). If ζ is known, Equation 7.34 can be used to directly estimate γ for a given synonymous codon set (averaged over genes).

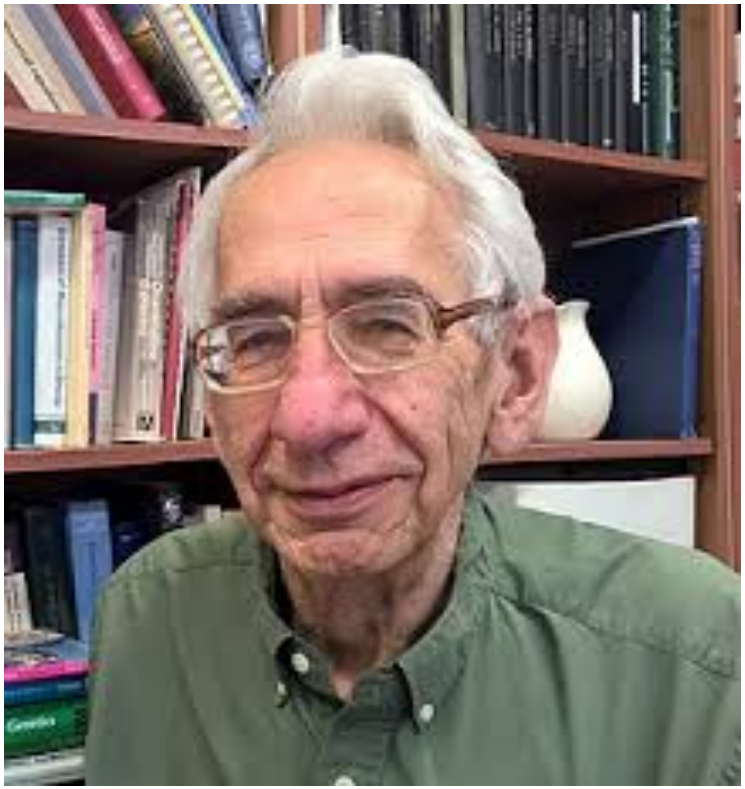
$4N_es \sim 1$ for Drosophila, suggesting bias might change over different genomic regions, as N_e changes

Drosophila codon bias varies over the genome

- Less extreme in
  - Regions of low recombination
    - More BGS, RS and hence lower N_e
  - In genes that are rapidly diverging
    - More RS and hence lower N_e
  - For long genes and in the middle of long exons
    - Much more interesting
- Key is that these differences are subtle and only apparent when a large number of sites are used (a genome-wide analysis)
- All consistent with selection at linked sites (RS or BGS) being important in shaping the genome

Fine-scale differences in bias

- What's behind very short-range effects (long vs. short genes, middle of long exons)
- Hill-Robertson effect:
  - Reduction in N_e due to selection at linked sites
- Small-scale HR effects
  - If multiple selected alleles are segregating, these can interfere with the effectiveness of selection, further weakening selection
  - Long exons: more regions that could be segregating sites under selection



Bill (W. G.) Hill

Alan Robertson